Employee Attrition Prediction:

This project works towards addressing the challenges of predicting employee attrition using the IBM HR Analytics Employee Attrition dataset.

As an HR professional working in the field of People Analytics, I am aware that a dashboard such as this would be extremely useful as employee attrition has profound implications across the organization. My goal was to build a predictive model that can identify employees at risk of leaving and, identify main features that are the drivers of attrition. So, through this project, I am not only attempting to predict employee attrition but also, provide insights into the underlying factors.

Methodology:

I began with data exploration to ensure that the data is clean, has no missing values and, review the summary statistics. After then, I proceeded to data preprocessing to ensure that all categorical features are encoded and carried out standardization of the numerical features using standard scaler. After that, I split the dataset 70:30 for training and test.

Feature selection was based on correlation analysis and then we evaluated many classification algorithms: Logistic Regression, Naive Bayes, Decision Tree, Random Forest, Support Vector Machines (SVM) with various kernels, and K-Nearest Neighbors (KNN). Hyperparameter tuning for KNN was performed using a validation curve to identify the optimal number of neighbors.

5-fold cross validation was conducted next to assess model performance on the key metrics of accuracy, precision, recall, and F1-score. Then Random Forest was caried out to rank the most important features.

Sigmoid was our best model with the highest accuracy score and a balanced performance across all metrics. And the top features affecting attrition where Age, Monthly income, Distance from Home, Years at company.

It seems that younger employees with lower monthly income are mote likely to leave. And less experienced employees with shorter tenure are more likely to leave than longer tenured employees. Inherently, commute is not a strong influencer to attrition. It is no surprise that lower job satisfaction and work-life balance are associated with higher attrition. Performance rating is also not a strong influencer to attrition.

Summary: Attrition is higher among younger, less experienced, lower-paid employees, and those with lower job satisfaction and work-life balance. These are key areas to focus on for retention strategies.

References

- IBM HR Analytics Employee Attrition & Performance Dataset. Kaggle
- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.

Appendix: Figures

All graphs and visualizations referenced in the analysis are included in the following appendix: